A Scalable Bootstrap for Massive Data

Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar and Michael I. Jordan

Discussion led by XianXing Zhang

January 17, 2014
Outline of today’s discussion

- Bootstrap
- Parametric Bootstrap and Bayesian Inference
- Bag of Little Bootstraps (BLB)
Why Care about Bootstrap

Assessing the quality of estimates (e.g., model parameters of interest) based upon training data

- It’s useful to quantify the uncertainty in that estimate.

- Such quality assessments provide more information than a simple point estimate.

- Model selection, smoothing, bias correction, active learning, among many potential uses
(Frequentist) Setting and Notation

- Assume sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, here $P \in \mathcal{P}$ is the true but unknown underlying distribution.

- Based on the sample we obtain an estimate $\hat{\theta}_n = \theta(\mathbb{P}_n) \in \Theta$, where $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ is the empirical distribution.

- The true (unknown) population value to be estimated is $\theta(P)$.

- Define $Q_n(P) \in \mathcal{Q}$ as the true underlying distribution of $\hat{\theta}_n$.

- The end goal is to compute some metric $\xi(Q_n(P))$, which summarizes $Q_n(P)$, e.g., $\xi$ might compute a confidence region, standard error, or a bias.
Bootstrap (Efron, 1979)

Basic idea

- Use the data-driven plugin approximation
  \[ \xi(Q_n(P)) \approx \xi(Q_n(P_n)) \]

- \( \xi(Q_n(P_n)) \) can be computed using Monte Carlo approximation:
  1. Repeatedly resample \( n \) points i.i.d. from \( P_n \).
  2. Compute the estimate on each resample.
  3. Form the empirical distribution \( Q_n \) of the computed estimates.
  4. Approximate \( \xi(Q_n(P)) \approx \xi(Q_n) \).

- Conceptually simple and powerful.
Bayesian Inference and the Parametric Bootstrap (Efron, 2012)

Why study this

- Parametric bootstrap has good computational properties when applicable.
- Makes connection between Bayes and frequentist inference transparent.

The bootstrap and Markov chain Monte Carlo

- Both are general-purpose methods for assessing statistical accuracy.
- Both methodologies operate in competing inferential realms: frequentist for the bootstrap, Bayesian for MCMC.
Basic idea

- Assume sample $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p(X|\theta)$, where the true underlying distribution is parameterized by some unknown parameter $\theta$.

- Estimate $\hat{\theta}$ by MLE.

- Draw bootstrap samples $X^*_b = X^*_b_1, \ldots, X^*_b_n$ i.i.d. from $p(X|\hat{\theta})$.

- Bootstrap replication $\hat{\theta}^*_b$ is computed from each sample $X^*_b$.

- The bootstrap replications can be thought as samples from the bootstrap density $f_{\hat{\theta}}(\hat{\theta}^*)$. 
Example

- $X_i \sim \mathcal{N}(\theta, 1)$.
- MLE $\hat{\theta} = n^{-1} \sum_{i=1}^{n} X_i$, and $\hat{\theta} \sim f_{\hat{\theta}}(\hat{\theta}) = \mathcal{N}(\theta, 1/n)$.
- A parametric bootstrap replication $\hat{\theta}^* \sim f_{\hat{\theta}^*}(\hat{\theta}^*) = \mathcal{N}(\hat{\theta}, 1/n)$.
- Here $\hat{\theta}$ is fixed at its observed value, while $\hat{\theta}^*$ is a random variable, and $f_{\hat{\theta}}(\hat{\theta}^*)$ is the bootstrap density.
- In the following $\hat{\theta}^*$ will be denoted as $\theta$ to harmonize notation with that for Bayesian inference.
Bayes Posterior Analysis

Given prior $\pi(\theta)$ and the likelihood function $f_\theta(\hat{\theta})$, the posterior expectation for $t(\theta)$:

$$E[t(\theta)|\hat{\theta}] = \frac{\int t(\theta)\pi(\theta)f_\theta(\hat{\theta})d\theta}{\int \pi(\theta)f_\theta(\hat{\theta})d\theta}.$$ 

Define the conversion factor, which is the ratio of likelihood to bootstrap density:

$$R(\theta) = \frac{f_\theta(\hat{\theta})}{f_\hat{\theta}(\theta)}.$$ 

Bootstrap integrals:

$$E[t(\theta)|\hat{\theta}] = \frac{\int t(\theta)\pi(\theta)R(\theta)f_\hat{\theta}(\theta)d\theta}{\int \pi(\theta)R(\theta)f_\hat{\theta}(\theta)d\theta}.$$
Bootstrap Estimation of Bayes Expectation $\mathbb{E}[t(\theta)|\hat{\theta}]$

- Recall parametric bootstrap replications $f_{\hat{\theta}}(\theta) \to \theta_1, \ldots, \theta_B$.

- Denote $t_i = t(\theta_i)$, $\pi_i = \pi(\theta_i)$, $R_i = R(\theta_i)$:

  $$\hat{\mathbb{E}}[t(\theta)|\hat{\theta}] = \frac{\sum_{i=1}^{B} t_i \pi_i R_i}{\sum_{i=1}^{B} \pi_i R_i}.$$  

- **Reweighting** Weight $\pi_i R_i$ on $\theta_i$.

- Importance sampling estimate:

  $$\hat{\mathbb{E}}[t(\theta)|\hat{\theta}] \to \mathbb{E}[t(\theta)|\hat{\theta}], \quad \text{as } B \to \infty.$$  

- Details on computing $R(\theta)$ under the exponential family assumption can be found at Efron's paper.
Computational Issues of Bootstrap

- Requires repeated estimator computation on resamples having size comparable to that of the original dataset.
- Alternatives (e.g., m out of n bootstrap and subsampling) are sensitive to tuning parameters.
Bag of Little Bootstraps (BLB)

- Partition the data into $s$ subsets, $I_1, \ldots, I_s$, each has size $|I_j| = b$, $\forall b$, and let $P_{n,b}^{(j)} = b^{-1} \sum_{i \in I_j} \delta x_i$ be the empirical distribution of subset $j$.

- For each subset, repeatedly resample $n$ points i.i.d. from $P_{n,b}^{(j)}$, then compute the estimate on each resample, form the empirical distribution $Q_{n,j}^*$ of the computed estimates, and approximate $\xi(Q_{n}(P_{n,b}^{(j)})) \approx \xi(Q_{n,j}^*)$.

- Then the BLB's estimate of $\xi(Q_{n}(P))$ is given by $s^{-1} \sum_{j=1}^{s} \xi(Q_{n}(P_{n,b}^{(j)}))$. 
Bag of Little Bootstraps (BLB)

Statistical correctness of BLB

- BLB has statistical properties that are identical to those of the bootstrap, details can be found in the paper.

Remarks

- BLB only requires repeated computation on small subsets of the original dataset ($b < n$).
- BLB permits computation on multiple subsamples and resamples simultaneously in parallel.
- Well suited to large-scale data and modern parallel and distributed computing architectures.